Comp. & Maths. with Appls. Vol. 12A, No. 2, pp. 235–244, 1986
Printed in Great Britain.

0097-4943/86 $3.00 + .00

CHOICE OF THE SMOOTHING PARAMETER AND EFFICIENCY OF k-NEAREST NEIGHBOR CLASSIFICATION

GREGORY G. ENAS
Lilly Research Laboratories, Indianapolis, IN 46285, U.S.A.

and

SUNG C. CHOI
Department of Biostatistics, Medical College of Virginia, Richmond, VA 23298, U.S.A.

Abstract—A simulation study was performed to investigate the sensitivity of the k-nearest neighbor (NN_k) rule of classification to the choice of k. The optimal choice of k was found to be a function of the dimension of the sample space, the size of the space, the covariance structure and the sample proportions. The nearest neighbor rules chosen using the k suggested by the simulations had correct classification rates at least as high as those rates for the linear discriminant function and the logistic regression method. In particular, the rule became more efficient as the difference in the covariance matrices increased, and also when the difference in sample proportion was large. An adaptive rule which selects k by iteratively maximizing the local Mahalanobis distance is shown to be efficient, thus abrogating the need to know the underlying population variance–covariance structure.

1. INTRODUCTION AND PRELIMINARIES

Fix and Hodges[6] introduced a novel approach to nonparametric classification by relying on the "distance" between points or distributions. The basic idea is to classify an individual to the population whose sample contains the majority of "nearest neighbors." Of the class of k-nearest neighbor rules, denoted by NN_k, the 1-nearest neighbor rule, where the observation is classified to the population from which the nearest neighbor is derived, has received special attention, notably by Cover and Hart[2], Cover[1] and Devroye[3].

Consider the problem of classifying a new observation into one of two known populations denoted by Π_i, $i = 1, 2$. First, we define a suitable metric over the sample space (e.g. Euclidean) and find the observation, whose population label is known from the training set T_R, which is "closest" to the new observation. Consider the rule which classifies the new observation to the population with the closest observation $\mathbf{x}$. In essence, this rule is based on the density or probability mass in a small neighborhood about the new observation $\mathbf{y}$. For example, the proportion of the N_1 observations from Π_1 which lie in a small hypersphere ΔN_1 containing $\mathbf{y}$ may be used to estimate the probability mass $p_1(\Delta N_1)$ in that neighborhood. Let the volume measure of the neighborhood be denoted by $V_1(\Delta N_1)$ so that the ratio $p_1(\Delta N_1)/V_1(\Delta N_1)$ becomes an estimate of the average value of $f_1(\mathbf{x})$ near $\mathbf{y}$. Under mild assumptions about the smoothness of $f_1(\mathbf{x})$ (i.e. continuity), this ratio is an estimate of $f_1(\mathbf{y})$. To obtain consistency, let the neighborhood about $\mathbf{y}$ shrink down to $\mathbf{y}$ as $N_1 \to \infty$, so that the average of $f_1(\mathbf{x})$ over the neighborhood will approach $f_1(\mathbf{y})$.

While these estimates are intuitively simple, they may not be practical when large samples are not available. The major difficulty lies in the choice of the sample size of the neighborhood ΔN_1. If this region is too small, the k sample points lying in that hypersphere will be too few to make an accurate estimate of the probability $p_1(\Delta N_1)$, given by k_1/N_1. On the other hand, if the region is too large, the proportion k_1/N_1 will not be a good approximation of $f_1(\mathbf{y})V_1(\mathbf{y})$. These two competing problems necessitate a compromise between either incurring an estimate with large variance or an estimate which is biased.

Fix and Hodges suggest that the p-dimensional sample space be transformed to a one-dimensional sample space by utilization of a suitable transformation which satisfies certain regularity assumptions. For example, in the two-population problem, a metric which satisfies these conditions is the Euclidean distance measure given by

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} (x_j - y_j)^2, \tag{1.1}$$

where x_j and y_j denote the jth component of $\mathbf{x}$ and $\mathbf{y}$, respectively.

The problem of evaluating $f_1(\mathbf{x})/f_2(\mathbf{x})$ can be then replaced by $g_1(d)/g_2(d)$ where $g_i(d)$, $i = 1, 2$ is the density function of $d(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}$ in Π_i. A consistent rule is constructed by taking the pooled sample of k observations from Π_i, $i = 1, 2$, which lie in a single region about zero. We let the total subset of k points to be chosen in such a way that $k \to \infty$ and $k/N \to 0$ as $N \to \infty$. If v equals the kth smallest value of the combined sample $d(\mathbf{x}, \mathbf{y})$'s, then

$$\hat{f}_{iv}(0) = k_i/(N_i v) \tag{1.2}$$

is a consistent estimate of $f_{iv}(0)$, $i = 1, 2$. The classification region R_i for Π_i is then R_1: $k_1/N_1 \geq ck_2/N_2$, R_2: otherwise, where c is a given positive constant chosen to reflect the relative importance of the two types of possible errors.

Cover and Hart showed that the large sample risk of the nearest neighbor rule with $k = 1$ is less than twice the optimal Bayes risk. If $T(\mathrm{NN}_k)$ denotes the probability of error for the k-nearest neighbor rule and $T(R^*)$ is the Bayes probability of error, then asymptotic error bounds for NN_k with $k = 1$ are given by

$$T(R^*) \leq T(\mathrm{NN}_k) \leq 2T(R^*)[1 - T(R^*)] \tag{1.3}$$

Cover and Hart also extended the NN_k rule to explicitly incorporate prior probabilities and generalized the rule to $m > 2$ populations. Devroye[3] extended the results of Cover and Hart to include all underlying distributions, not just continuous densities. Thus the bounds of Eq. (1.3) hold when the underlying joint distribution is a mixture of continuous and discrete distributions. Devroye also showed that

$$\lim_{N \to \infty} \mathrm{pr}|T(\mathrm{NN}_k) - 2T(R^*)[1 - T(R^*)]| \to 0 \tag{1.4}$$

when ties between observations are broken satisfactorily. Thus the efficiency of the NN_k rule decreases as the sample size of the training set increases. Devroye[4] has also shown further results pertaining to the asymptotic probability of error bound given in Eqs. (1.3) and (1.4). Rewriting Eq. (1.3), Devroye showed that $\lim_{N \to \infty} \sup T(NN_k) \leq c_k T(R^*)$, where c_k is a sequence of numbers of the form, $(1 + (\alpha/k))(1 + 0(1))$ as $k \to \infty$ for some fixed α.

2. THE OPTIMAL CHOICE OF k

2.1 *Optimal* k *for small samples*

The very attractive error bound (1.3) holds only if k is chosen such that $k \to \infty$ and $k/N \to 0$ as $N \to \infty$. Hence it is very important to answer the question dealing with the optimal choice of k for the small to moderate sample sizes found in practice. First, consider a k-nearest neighbor density estimate of $f_i(\mathbf{y})$, $i = 1, 2$. Fukunaga and Hostetler[8] have developed a functional form for the optimum k in terms of N_i, the dimensionality of the sample space denoted as p, and the underlying probability distribution, by expanding that underlying distribution in a Taylor series about the test point $\mathbf{y}$. They derive an approximation to the volume/coverage relationship of the underlying distribution in the neighborhood of $\mathbf{y}$. Minimization of an approximation to the mean-square criterion, given by

$$J_i(\mathbf{x}) = E[(f_i(\mathbf{x}) - \hat{f}_i(\mathbf{x}))^2],$$

or the integral mean-square error criterion

$$I_i = \int J_i(\mathbf{x})\, d\mathbf{x} \tag{2.1}$$

results in the optimal choice of k and, in the case of Eq. (2.1), the optimal k being independent of the new observation $\mathbf{y}$.

They also show the optimum matrix V which minimizes Eq. (2.1) for use in a metric with

quadratic form, such as the Euclidean metric, is the inverse covariance matrix of the underlying distribution. This property will hold for any distribution within a general class of distributions that can be made circularly symmetric by a linear transformation. Hence, when the underlying distribution is normal, the inverse covariance matrix so often used is indeed optimal. A true distribution-free result for optimal k will not be derived, for as Fisher[5] showed, the choice of k is a function of the smoothness of the underlying distribution.

2.2 *Empirical investigation of NN_k rules*

How should the number of nearest neighbors be chosen such that the probability of misclassification is minimized for small and moderate sample sizes? Fix and Hodges[7] obtained exact and asymptotic expressions for the probabilities of misclassification of the NN_k rule assuming random samples of equal size N_i from each population Π_i, $i = 1, 2$, and an underlying univariate normal distribution with equal variances. It seems clear, however, that the optimal choice of k depends not only on the size of the sample but also on the covariance structures within each population and the sample proportions for each population within the total sample. There appears to be no previous study dealing with these issues. We wish to study the effects of different covariance matrices and sample sizes on the probabilities of misclassification for certain choices of k. We assume equal *a priori* probabilities.

The simulation study is based on observations of the form $(X_1, X_2, X_3, X_4, X_5 \mid \Pi_i)$, $i = 1, 2$ where X_1 and X_2 have a bivariate normal distribution, and X_3 has a Bernoulli distribution. We assume a trichotomous variable Z is represented by X_4 and X_5 as follows: $X_4 = 0$ and $X_5 = 0$ if $Z = 0$; $X_4 = 1$ and $X_5 = 0$ if $Z = 1$; $X_4 = 0$ and $X_5 = 1$ if $Z = 2$. The joint distribution of (X_3, Z) for each population is held constant as follows for Π_i, $i = 1, 2$:

Π_1		Z			Π_2		Z		
		0	1	2			0	1	2
X_3	0	0.05	0.10	0.10	X_3	0	0.05	0.40	0.25
	1	0.10	0.15	0.50		1	0.15	0.10	0.05

The following three different covariance matrices were considered:

Matrix 1

X_1	2.00				
X_2	1.50	2.00			
X_3	0.20	0.10	0.20		
X_4	−0.20	−0.20	−0.05	0.20	
X_5	0.50	0.50	−0.05	0.15	0.20

Matrix 2

X_1	1.75				
X_2	0.67	1.75			
X_3	−0.10	−0.10	0.20		
X_4	0.20	0.20	−0.05	0.25	
X_5	0.20	−0.20	−0.05	−0.15	0.20

Matrix 3

X_1	1.80				
X_2	0.50	1.80			
X_3	0.05	0.05	0.20		
X_4	0.10	0.10	−0.05	0.25	
X_5	−0.35	−0.35	−0.05	−0.15	0.20

The nearest neighbor rule we chose to evaluate was the original NN_k rule proposed by Fix

and Hodges. The choice of k for optimal classification was investigated by considering $k = [N^{j/8}]$ for $j = 0, 2, 3, 4$ and 5, where $[t]$ denotes the closest odd integer to t and N is the total pooled sample size of the training set T_R. The following three cases were defined from different combinations of the covariance matrices:

	Matrix	
Case	Π_1	Π_2
A	1	1
B	1	2
C	1	3

Note that case A allows investigations of performance under the assumption of equal covariance matrices between populations. Cases B and C allow consideration of gradually increasing disparity between population covariance structures.

In the simulation study, pooled sample sizes of $N = [50, 134, 354]$ were considered for the training set T_R and the size of the test set T_E was held fixed at 50. Different sample mixing proportions of both T_R and T_E were considered to reflect the different proportions of sample sizes in $\Pi_1 : \Pi_2$, respectively. These ratios are 1:4, 1:2, 1:1, 2:1 and 4:1.

Based on T_E, unbiased estimates of correct classification rate (CCR) of the NN_k rule constructed from T_R for various cases were tabulated. Tables 1–3 present the results of the study showing the effect of sample proportions, sample size and covariance structure on k. From these results the following observations may be made concerning the behavior of NN_k with respect to k.

1. As within population covariance matrixes become more dissimilar, NN_k rules tend to be more efficient, that is, CCR appears to be greater. Simultaneously, CCR decreases for relatively larger values of k as covariance matrixes become more dissimilar.
2. Optimal k decreases as within population correlation structures become more different as in case C.
3. Optimal k tends to fluctuate more with differing sample sizes when within population covariance matrices are unequal (cases B and C) than for equal covariance matrices (case A).
4. All rules tend to improve as total sample size N of T_R increases yet the $k = 1$ rule becomes increasingly inefficient for larger samples relative to small samples.
5. CCR increases as the difference in sample proportions increases.
6. The functional form of our estimate of k seems to be a reasonable approximation as the behavior of the power function remains rather constant within each covariance and sample proportion structure.

As seen from Tables 1–3, the actual correct classification rates, CCR, seem to follow rather smooth function of the choice of k. Optional choices for k may be made for each particular situation.

2.3 *Comparison with other methods*

We now turn our attention to the next question of interest, namely how efficient is the optional NN_k rule relative to the linear logistic classification model (LOG) and the linear discriminant classification model (LDF) under the same conditions described above. These latter two models have been shown to be sometimes robust with respect to non-normality, the logistic model especially suited to handle mixtures of continuous and discrete variables.

Simulation experiments described in Section 2.2 were performed to compare the three methods. Table 4 shows the optimal value of CCR for the NN_k rule compared to the CCR for the LOG and the LDF.

Interactive solutions were found for the parameters of the logistic functions in most cases where starting values were always set equal to zero. However, in cells marked by an $*$, convergence was obtained in less than 10% of the 100 trials per cell. This occurred only twice. Convergence was obtained in cell trials for most of the larger sample sizes.

Table 1. Correct classification rate (%) for different values of k for covariance structure A

N	Sample Proportion	$k = N^{i/8}$				
		i = 1	i = 2	i = 3	i = 4	i = 5
50	1:1	75.3	76.7	77.2	76.5	76.6
	2:1	76.4	79.3	78.9	78.2	77.3
	4:1	81.5	83.1	83.2	82.6	82.1
134	1:1	77.7	80.8	81.0	80.9	80.4
	2:1	77.5	80.0	81.0	80.3	78.1
	4:1	83.4	84.8	84.7	84.5	83.3
355	1:1	77.2	82.1	82.2	81.5	80.5
	2:1	79.4	82.5	82.9	82.3	80.9
	4:1	82.9	86.4	86.6	86.4	84.4

Table 2. Correct classification rate (%) for different values of k for covariance structure B

N	Sample Proportion	$k = N^{i/8}$				
		i = 1	i = 2	i = 3	i = 4	i = 5
50	1:4	85.4	84.0	82.7	81.4	79.6
	1:2	81.5	80.0	77.1	74.0	71.0
	1:1	76.3	78.5	77.7	77.0	74.2
	2:1	80.8	80.8	80.0	79.0	77.6
	4:1	85.9	87.1	85.7	84.1	81.4
134	1:4	86.7	87.1	86.0	84.5	80.9
	1:2	84.3	84.4	83.1	80.3	84.7
	1:1	80.2	81.9	82.2	80.9	79.1
	2:1	81.4	82.6	83.2	82.7	79.1
	4:1	86.1	88.4	88.1	87.3	84.5
354	1:4	87.4	88.4	88.0	88.1	83.5
	1:2	85.0	86.1	85.6	83.8	79.9
	1:1	80.0	83.6	84.1	83.6	82.1
	2:1	82.0	84.9	84.9	85.4	82.4
	4:1	85.5	88.8	89.0	88.6	86.7

Table 3. Correct classification rate (%) for different values of k for covariance structure C

N	Sample Proportion	$k = N^{i/8}$				
		i = 1	i = 2	i = 3	i = 4	i = 5
50	1:4	85.0	84.2	82.2	81.0	79.7
	1:2	80.2	78.5	76.4	74.3	70.1
	1:1	84.1	82.3	80.0	77.9	73.0
	2:1	84.2	83.3	80.7	78.3	74.8
	4:1	86.8	85.6	83.7	83.0	80.8
134	1:4	86.7	87.6	86.8	85.2	80.7
	1:2	82.9	83.6	82.8	81.9	75.4
	1:1	86.2	86.8	86.9	85.0	81.4
	2:1	86.4	86.9	86.7	84.4	78.2
	4:1	89.1	89.6	89.3	87.5	82.8
354	1:4	87.6	88.9	89.0	88.3	82.0
	1:2	82.9	84.7	84.7	83.9	81.0
	1:1	87.0	88.7	87.9	86.7	83.6
	2:1	87.6	88.8	89.0	89.4	89.4
	4:1	88.6	90.4	80.6	90.7	90.5

Table 4. Correct classification rates (%) for LDF, LOG, and optimal NN_K with respect to sample size, sample proportion and covariance structure

Sample Ratio	Sample Size	Covariance Structure								
		A			B			C		
N_1/N_2	$N_1+N_2=N$	NN_K	LOG	LDF	NN_K	LOG	LDF	NN_K	LOG	LDF
0.20	50	----	----	----	85.0	82.0*	77.0	85.4	80.2**	76.7
	134	----	----	----	87.6	82.2**	76.6	87.1	81.3	76.4
	354	----	----	----	89.0	83.2**	76.5	88.4	82.0	78.4
0.33	50	----	----	----	80.2	70.5**	73.9	81.5	78.8**	74.0
	134	----	----	----	83.6	73.8*	75.7	84.4	77.6	78.4
	354	----	----	----	84.7	72.9**	76.4	86.1	77.7	79.8
0.50	50	77.2	80.7**	80.5	78.5	77.0**	77.3	84.1	78.1	73.2
	134	81.0	83.7	83.7	82.2	77.8	78.6	86.9	76.0	75.6
	354	82.2	84.1	83.7	84.1	77.6	77.9	88.7	75.1	74.6
0.67	50	79.3	82.8**	81.9	80.8	81.1	79.2	84.2	76.7**	74.9
	134	81.0	83.7	81.8	83.2	81.5	79.2	86.9	76.5	74.5
	354	82.9	84.1	82.3	85.4	83.0	80.5	89.4	78.2	76.0
0.80	50	83.2	85.1**	80.2	87.1	87.7**	82.0	86.8	83.4**	76.1
	134	84.8	87.0	82.5	88.4	88.4	83.3	89.6	85.9	77.4
	354	86.6	87.4	82.9	89.0	88.8	83.0	90.7	87.1	77.8

* = $\leq$ 10 observations; ** = 50 $\leq$ m < 100 observations

As expected, the LOG and LDF perform better than NN_k when the population covariance matrices are equal, namely for case A. However, these differences are not pronounced. On the other hand, for case A the LDF suffers as sample proportions become dissimilar while NN_k performs increasingly more efficiently, as does the LOG. The NN_k is at least as efficient and generally more efficient than the LOG for small samples as the covariance matrices begin to exhibit slight dissimilarity as in case B. It is interesting to note that the NN_k is uniformly more efficient for large sample sizes under case B. Note that for each proportion, NN_k appears to be more efficient than the LOG and LDF when the covariance matrices are very much different as in case C. As the covariance matrices become more unequal this efficiency appears to become more pronounced for all sample sizes. It is interesting to note that even where the covariance matrices are equal the LDF is relatively less efficient when there are large differences in sample proportions than for equal sample proportions. As covariance matrices become more different, the LDF performs better for larger differences in sample proportions. The LOG performs more efficiently as the difference in sample proportions increases for any given covariance structure.

However, for the mixtures of variables considered, the LOG performs more efficiently than the LDF when sample proportions exhibit large differences and covariance matrices are unequal. Yet the most significant finding is the dominating performance of the NN_k over the LOG for situations considered under covariance cases B and C. The NN_k rule performs best when covariance matrices are quite different and differences in sample proportions are large. Hence it is encouraging to note the surprising efficiency of the NN_k for all types of covariance structures. Even though the NN_k utilizes a pooled sample covariance matrix in determining the standardized Euclidean distance from test point to the points in the training set, its ability to construct a locally efficient rule even under global covariance inequality is advantageous. The distance measure utilized seems to scale each variable appropriately and mixtures of variables are treated adequately.

3. AN ADAPTIVE NEAREST NEIGHBOR RULE

3.1 *Adaptive nearest neighbor* (ANN) *rule*

Patrick[10] summarizes another class of nearest neighbor rules which we will denote as ANN rules. These rules generate a sequence of regions of hyperspheres from each population instead of utilizing just one sequence of regions containing the pooled samples from all populations. These rules are equivalent to the NN_k rules yet enjoy some distinct advantages, including utilization of an even number for k as well as explicit utilization of prior probabilities. Goldstein[9] suggests an ANN rule in which samples of equal size $N = N_i$, $i = 1, 2$, be taken from each population Π_i. These samples are then ordered separately, with respect to a Euclidean metric, in ascending order. Goldstein suggested subsequent allocation be made to the population whose k_ith observation, $i = 1, 2$, was closest. Rabiner *et al.*[11] modified this procedure to utilize greater information from the observations located within the hypersphere and assigning the observation to the population whose centroid is closest.

The idea is to exploit the ANN formulation to obtain maximal local separation instead of seeking maximal global separation. We seek to find those optimal regions containing $\mathbf{y}$ for which the linear function of explanatory variables constructed from only those observations within the regions has maximum "variance between regions" relative to the "variance within regions." Let $V_i(\mathbf{x})$, $i = 1, 2$, be the volume of the region defined by the nearest neighbors from Π_i. The problem is to find the volumes which have maximal separation while maintaining constant coverage over the regions. Assignment is made to that population for which the weighted average distance of the observations within their region to $\mathbf{y}$ which is to be classified is smallest.

The problem is to construct these regions in some "optimal" fashion. Among many possibilities, one ready solution to this problem is to incorporate the Mahalanobis distance measure. Given a new observation $\mathbf{y}$, we propose to choose $p^* = p_i$, $i = 1, 2$, such that the following attains a local maximum:

$$\Delta_k^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_k^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

where $\bar{\mathbf{x}}_i$, $i = 1, 2$, are the two mean vectors, and $\mathbf{S}_k$ denotes the "local" sample pooled

covariance matrix for the neighborhood each based on k-nearest neighbors to $\mathbf{y}$. If the local maximum is for p such that $k_1 + k_2$ is greater than $N^{-1/2}$, then selection of nearest neighbors is terminated, and $p^* = N^{-1/2}$.

It may happen that the observations selected in the initial stages of selection are very close to each other, and the dispersion in the given region is very small. Hence $\mathbf{S}_k$ will be nearly singular in these cases and Δ_k^2 will not be explicitly utilized until the volume $V_i(\mathbf{x})$ of the region is relatively large, thus sufficient for nonsingular $\mathbf{S}_k$.

Thus, as the sample sizes of each sample increase, the number of selected observations will have to increase in order that $\mathbf{S}_k$ be nonsingular. Thus the condition for optimality that $k \to \infty$ and $k/N \to 0$ as $N \to \infty$ will be met. Even when $\mathbf{S}_k$ is nonsingular, a local maximum for Δ_k^2 may not be obtained. This may include cases when $\mathbf{y}$ lies in the outlying region of the sample space, or when a majority of components of $\mathbf{X}$ are discrete. Then the upper bound $N^{-1/2}$ for p^* is obtained. In practice, this arbitrary bound has led to efficient rules. (A copy of the adaptive nearest neighbor algorithm (ALGO) which solves the above problem is available from the first author.)

3.2 *Empirical results with adaptive algorithm*

We desire to compare the adaptive nearest neighbor technique based on ALGO proposed in Sec. 3.1 with the NN_k techniques and the linear discriminant function. To put the adaptive technique to its most stringent test involving the same distributional characterizations described in Sec. 2, we choose to apply the technique to training sets of size 50, the observations being generated in equal proportions from both populations, characterized by the same covariance matrices of case A. The ANN rule given by ALGO was compared with NN_5, NN_7 and LDF based on a small simulation study. The mean correct classification rate (CCR) and the standard error on the basis of 50 observations in the test set over 100 independent replications is given as follows:

ALGO	NN_5	NN_7	LDF
79.5 ± 6.5	78.1 ± 7.1	77.8 ± 6.5	80.7 ± 6.0

The optimal NN_k rules used in this situation were established using the guidelines in Sec. 2. As might be expected, the LDF slightly outperforms the nearest neighbor rules under these given conditions of identical covariance matrices and equal sample sizes. Yet the more interesting finding is the slightly better performance of the algorithm compared to the optimal NN_k rules.

The adaptive technique has also been shown to induce a balance between the conditional probabilities of misclassification $p(1 \mid 2)$ and $p(2 \mid 1)$. This would be desired if the sample proportions were quite unequal. We would expect the NN_k rules as well as the LDF to be biased towards the population represented by the larger sample. This bias is likely to be deflated with the adaptive procedure, as it gives increased weight to those observations from the small sample. So, to investigate the effects of unequal sample proportion as well as differing convariance structure, we again simulated the conditions described in Sec. 2 for the two multivariate populations consisting of a mixture of two continuous and three discrete variables and chose optimal NN_k rules using the guidelines. Here we chose the sample ratio to be 4:1 from populations characterized by covariance structure given by case C. Random samples of 273 and 71 were taken from Π_1 and Π_2, respectively, from which the different classification rules were constructed under assumptions of equal prior probabilities. Then a test set including samples of 40 and 10 from Π_1 and Π_2, respectively, was subsequently classified by each rule.

Due to the unbalanced nature of the problem, a utility function defined as $(p(1 \mid 1) + p(2 \mid 2)$ was calculated. The mean utility and the standard error for each procedure based on 100 independent replications are as follows:

ALGO	NN_9	NN_{19}	LDF
1.76 ± 0.01	1.70 ± 0.01	1.60 ± 0.02	1.51 ± 0.02

Note the apparent superiority of the ALGO rule over the optimal NN_k rules and the LDF in this situation. We expected the LDF to be least efficient in this case.

4. CONCLUDING REMARKS

For a mixture of categorical and continuous variables, a class of nonparametric classification rules has been shown empirically to perform as well or better than the classical parametric methods for small to moderate sample sizes. The choice of k for optimal performance of the nonparametric nearest neighbor rules has been shown to be dependent on the sample proportions and the underlying covariance structure for small samples. Although further work is needed, our study suggests the following rough guidelines in selecting k for optimal classification based on the size of the training set N:

		Difference Between Sample Proportions	
		Small	Large
Difference Between Covariance Matrices	Small	$N^{3/8}$	$N^{2/8}$
	Large	$N^{2/8}$	$N^{3/8}$

It is not surprising to see the greater efficiency of the NN_k rules relative to LDF and logistic regression methods as the covariance matrices become more dissimilar, or the differences between the sample sizes and shapes of the distributions increase. The local optimality of the NN_k rules is seen dramatically when global dispersion for each population is grossly dissimilar.

The adaptive rule presented in Section 3 is a simple and intuitive attempt to choose the optimal size of k. An equal percentage of observations nearest **y** from each sample are taken to develop the classification rule for **y**. The optimal percentage p^* is obtained such that maximal separation between the region's centroids is realized relative to the pooled dispersion within the regions. The efficiency of this solution is dependent on the degree of local discrimination between the groups.

It is possible that the observations in both regions could be pooled and ranked in ascending order depending on their distance to **y**. The population from which the minimum sum of ranks was generated would be chosen for discrimination. Though distributional properties may possibly be worked out for these sums, the small sample efficiency of this procedure would intuitively be less than the procedure incorporating the original distances. The algorithm may be easily extended to facilitate classification into $m > 2$ populations.

Acknowledgements—This work was supported in part by NINCDS Grant No. NS-12587. The gracious and copious secretarial assistance of Mrs. Dhea Rigdon is fully appreciated.

REFERENCES

1. T. M. Cover, Estimation by the nearest-neighbor rule. *IEEE Trans. Inf. Theory* **II-14**, 50–55 (1968).
2. T. M. Cover and P. E. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inf.* Theory **IT-13**, 21–27 (1967).
3. L. Devroye, On the equality of Cover and Hart in nearest neighbor discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.* **3**, 75–78 (1981).
4. L. Devroye, On the asymptotic probability of error in nonparametric discrimination. *Ann. Stat.* **6**, 1320–1327 (1981).
5. F. P. Fisher, K-nearest neighbor rules. Ph.D. Dissertation, School of Electrical Engineering, Purdue University, Lafeyette, IN (1971).
6. E. Fix and J. L. Hodges, *Nonparametric Discrimination: Consistency Properties*. Project No. 21-49-004, Report No. 4, U.S. Air Force School of Aviation Medicine, Randolph Field, TX (1951).
7. E. Fix and J. L. Hodges, *Discriminatory Analysis, Small Sample Performance*. Project 21-49-004, Report No. 11, U.S. Air Force School of Aviation Medicine, Randolph Field, TX, August (1952).
8. K. Fukunaga and L. D. Hostetler, Optimization of k-nearest neighbor density estimates. *IEEE Trans. Inf. Theory* **IT-19**, 320–326 (1973).
9. M. Goldstein, K_n-nearest neighbor classification. *IEEE Trans. Inf. Theory* **IT-18**, 627–630 (1972).
10. E. A. Patrick, *Fundamentals of Pattern Recognition*. Prentice Hall, Englewood Cliffs, NJ (1972).
11. L. R. Rabiner, S. E. Levenson, A. E. Rosenberg and J. G. Wilson, Speaker-independent recognition of isolated words using clustering techniques. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-27**, 339–349 (1979).